PREDICTIVE SAMPLE REUSE APPROACHES TO MODEL SELECTION

by

Seymour Geisser* and William F. Eddy
University of Minnesota Carnegie-Mellon University

Technical Report No. 280

February, 1977

*Research partially supported by U.S. Army Grant No. DAHC04-74-G-0216.

Predictive Sample Reuse Approaches to Model Selection

by

Seymour Geisser
University of Minnesota

and

William F. Eddy
Carnegie-Mellon University

1. Introduction

Consider the problem of inferring which of several ostensible models $M_1, M_2, \ldots, M_m$ best explain a set of data $\underset{\sim}{x}^{(N)}$ based on N independent observations. One way, and in a sense the most useful, is to determine which of these models renders the "best" predictions for future observations. If the process generating observations was not subject to stochastic variation, or only trivially so, and critical conditions led to completely distinct predictions, it would be a simple matter to select the most appropriate model of those under consideration. If it were clear that the process was subject to random variation and the models were only specifiable up to a known distribution function $F_k(\cdot|M_k)$ then a ranking of the likelihoods under each alternative model and selection of the most likely model given the data, is possible. If prior probabilities reflecting the "truth" of the various models were assumed then the most probable of those considered is clearly calculable.

Problems confronting research workers and statisticians are often not quite so simple. At best the models are such that the distributions are specifiable only up to a set of unknown parameters. They may, in one set of circumstances, involve completely distinct distribution functions and parameters or in another be a nesting of models, e.g., under one model two

parameters may differ while for another the parameters may be the same. When many other assumptions are satisfied these problems can be handled within the Bayesian framework of hypothesis testing. But this has not really been a resolution acceptable to many mainly because of the necessary reliance on a host of subjective assumptions. In particular for the nesting situation it seems difficult to escape the awkward introduction of lumps of probability to avoid a zero posterior probability of a model which asserts the equality of several parameters relative to a model which assumes the contrary. Further difficulties ensue when reasonable distributional assumptions with regard to the process generating the data are not tenable. It may also occur that the models are only somewhat vaguely specified, e.g., one model may specify that a label is relevant for prediction while another asserts that it isn't.

In this paper Predictive Sample Reuse (PSR) techniques will be used to supply some partial resolutions for these problems. In a technical report by Lee and Geisser (1972) the method was utilized to determine, for particular sets of growth curve data, which of a wide variety of plausible approaches, each of which resulted in a method for predicting future observations from the data, would be most appropriate. For a particular set of data each alternative approach yielded a predictor for an omitted observation based on the rest of the data. This was repeated for each observation and various relevant summary measures of the discrepancy of the predicted values from the observed values were calculated. These summary measures then provided a basis for the selection of the method apparently most appropriate for the given set of data. This simple approach appeared to be particularly useful because the more than a dozen prediction methods ran the gamut of highly structured Bayesian paradigms to simple data analytic models.

In the next section we assume sampling distributions are specified but that the parameters are unknown and subjective prior distributions for them are unavailable. For this situation we propose a blending of sample reuse and quasi-Bayesian procedures as a solution. A quasi-Bayesian procedure is one that allows improper prior distributions and the use of a product of conditional predictive distributions as a criterion in place of the more logically appropriate joint predictive distribution. Applications and illustrations are presented for some particular cases.

In the third section we abandon distributional assumptions and present and discuss some simple data analytic sample reuse solutions.

2. High Structure Selection - General Setup

Given a set of data $\underset{\sim}{x}^{(N)} = (x_1,\ldots,x_N)$, arising from independently distributed random variables, suppose that for each x_j there is an associated set z_j, which incorporates all that is assumed known with certitude about x_j. Assume further that a number of possible models $M_1,\ldots,M_m$ could have generated the data or can tentatively offer a satisfactory explanation of the data. The models should imply distinct distributions for the data and may induce a partition of the data based on the known associated set $\underset{\sim}{z}^{(N)} = (z_1,\ldots,z_N)$. For example, a particular M_k may imply that there are two different populations π_1 and π_2 based on the fact that the data are capable of a recognizable partition $\underset{\sim}{x}^{(N)} = (\underset{\sim}{x}_1^{(N_1)}, \underset{\sim}{x}_2^{(N_2)})$ where $\underset{\sim}{x}_i^{(N_i)} = (x_{i1},\ldots,x_{iN_i})$. By this we mean that each z_j includes as one of its components a recognizable label denoted by $i = 1$ or 2; and that the $\underset{\sim}{x}_i^{(N_i)}$ could be viewed as a random sample from π_i. An alternative model $M_{k'}$ might posit that the labels 1 and 2 were irrelevant and that $\pi_1 = \pi_2 = \pi$. In either situation the distribution functions may be partially or completely specified. If the distributions were completely specified under each model then one would compute, assuming $\underset{\sim}{x}^{(N)} = (x_1,\ldots,x_N)$ was the realized value of a set of independent random variables $\underset{\sim}{X}^{(N)}$,

$$L_k = f(\underset{\sim}{x}^{(N)} | \underset{\sim}{z}^{(N)}, M_k) = \prod_{j=1}^{N} f_j(x_j | z_j, M_k)$$

the probability density or likelihood as a criterion of assessment. However the density $f(\underset{\sim}{x}^{(N)} | \underset{\sim}{z}^{(N)}, M_k)$ will, in most instances, be specified up to a set of unknown parameters $\underset{\sim}{\theta}_k$ as indicated by M_k. L_k can presumably be estimated by inserting for $\underset{\sim}{\theta}_k$ the m.l.e. $\hat{\underset{\sim}{\theta}}_k(\underset{\sim}{x}^{(N)})$ under M_k. In particular if $M_k \subseteq M_{k'}$, as will sometimes be the case, this is of little use for

direct comparisons because the maximization obviously requires that $L_{k'}(\hat{\underset{\sim}{\theta}}_{k'}) \leq L_k(\hat{\underset{\sim}{\theta}}_k)$. However the quotient of the two maximized likelihoods is the basis for Likelihood Ratio tests. Here the PSR method will be utilized so that direct comparisons will still be meaningful. Let $\underset{\sim}{x}_{(j)}^{(N-1)}$ represent the data set $\underset{\sim}{x}^{(N)}$ but with x_j omitted. Further consider a predicting density $f(x|\underset{\sim}{x}^{(N)}, M_k)$ that could be used to predict future observations when M_k is true. For example one could choose as a predicting density $f(x|\hat{\underset{\sim}{\theta}}_k, M_k)$ having the same form as the postulated sampling density $f(x|\underset{\sim}{\theta}, M_k)$ but with the m.l.e. estimator $\hat{\underset{\sim}{\theta}}_k$ substituted for $\underset{\sim}{\theta}_k$. We prefer, for the most part, an alternative choice for the predicting density--one that would be indicated by a predictive Bayesian analysis. We shall discuss this point subsequently.

Nonetheless once the choice is made the predicting density is modified so that it can be applied to the observations already in hand. This is accomplished by considering as predicting density for X_j, $f_j(x_j|\underset{\sim}{x}_{(j)}^{(N-1)}, z_j, M_k)$ and then computing

$$L_k = \prod_{i=1}^{N} f_j(x_j|x_{(j)}^{(N-1)}, z_j, M_k) , \quad k = 1,\ldots,m$$

and obtaining

$$L_k^* = \max_k L_k .$$

All other things being equal the model M_{k^*} is then selected as the most appropriate among those under consideration. It is worthwhile to note that the product L_k is formed by treating the X_j as though they were predictively

independent--which they are when conditioned on the parameters as reflected in their joint sampling distributions. However if the actual predictive densities derived from a Bayesian analysis had been used, the joint distribution of the X_j's would invariably be unconditionally (predictively) dependent, Geisser (1966). The product L_k then is to be regarded as a compromise between Bayesian and non-Bayesian methods in the following sense: Firstly, for the sake of computational convenience the product of the conditional densities of X_j given $\underset{\sim}{X}_{(j)}^{(N-1)} = \underset{\sim}{x}_{(j)}^{(N-1)}$, $j = 1,\ldots,N$, is used rather than the correct joint predictive density. Secondly we note that the conditional predictive density of X_j will depend on $\underset{\sim}{X}_{(j)}^{(N-1)} = \underset{\sim}{x}_{(j)}^{(n-1)}$ as well as the form and hyperparameters of the prior distribution of $\underset{\sim}{\theta}_k$ but that the joint predictive density depends only on the latter. This appears to us to put too much emphasis on the prior distribution. Thirdly if the prior distribution of $\underset{\sim}{\theta}_k$ is improper then the joint predictive density of $\underset{\sim}{X}^{(N)}$ will also be improper but we often find it convenient to use improper priors. In summary we are modifying a highly structured Bayesian procedure by laying greater stress on the data.

The use of the predictive density as either an estimate or a surrogate for the sampling density was suggested by Geisser (1971). From a Bayesian viewpoint this is a better procedure than utilizing as an estimate the original sampling density with the m.l.e. $\hat{\underset{\sim}{\theta}}_k$ substituted for $\underset{\sim}{\theta}_k$. Aitchison (1975) presents a frequentist justification for this fact using the Kullback-Leibler (1951) directed measure of divergence.

Recently M. Stone (1976) derived a result which is of some interest in elucidating the relationship between PSR and other methods. He showed

that if one used as predicting density the reused estimative sampling density, constructed by substituting the m.l.e. of the parameter set for the parameter set itself excluding the observation inserted into the density, then this reused quasi-likelihood criterion was asymptotically equivalent to a selection criterion proposed by Akaike (1973) which will be given below. In other words, if X_j has density $f_j(x_j|z_j,\underset{\sim}{\theta}_k,M_k)$ $j = 1,\ldots,N$, and the reused quasi-likelihood predicting density is

$$f_j(x_j|z_j,\hat{\underset{\sim}{\theta}}_{k(j)},M_k) \tag{2.1}$$

where $\hat{\underset{\sim}{\theta}}_{k(j)}$ is the m.l.e. of $\underset{\sim}{\theta}_k$ with x_j omitted, then

$$\hat{L}_k = \prod_{j=1}^{N} f_j(x_j|z_j,\hat{\underset{\sim}{\theta}}_{k(j)},M_k) \to e^{-p(M_k)} \prod_{j=1}^{N} f_j(x_j|z_j,\hat{\underset{\sim}{\theta}}_k,M_k) \tag{2.2}$$

as $N \to \infty$, where $p(M_k)$ is the number of unknown parameters of $\underset{\sim}{\theta}_k$ as specified by M_k. The r.h.s. of (2.2) is essentially the Akaike criterion for model selection in that the largest value, for $k = 1,\ldots,m$, indicates the choice of the model.

For two alternative models M_k and $M_{k'}$,

$$\lambda = \prod_{j=1}^{N} f_j(x_j|z_j,\hat{\theta}_k,M_k) \Big/ \prod_{j=1}^{N} f_j(x_j|z_j,\hat{\theta}_{k'},M_{k'}) , \tag{2.3}$$

is the likelihood ratio test criterion. Now, under fairly general conditions, $-2 \log \lambda$ tends asymptotically to a χ^2 with $p(M_{k'}) - p(M_k) = p$ degrees of freedom when M_k is the true model. Further, for deciding between two models M_k and $M_{k'}$, the reused quasi-likelihood predicting density procedure is equivalent to $-2 \log \dfrac{\hat{L}_k}{\hat{L}_{k'}} > 0$ for choosing $M_{k'}$ in favor of M_k.

Hence, α , the significance level of the criterion asymptotically tends to

$$P\{\chi_p^2 > 2p\} \ , \tag{2.4}$$

where χ_p^2 is a chi-squared variate with p degrees of freedom.

It can also be shown that the particular reused quasi-Bayes predicting densities set forth in the following subsections 2.1-2.3 have this same asymptotic property as the reused quasi-likelihood predicting density. However since we have not produced a general method for obtaining these quasi-Bayes predicting densities, we cannot give a general theorem, but a case by case verification is possible. As indicated previously, however, our choice of the quasi-Bayes predicting densities is predicated on the fact they provide better "estimates" of the sampling density than the m.l.e. estimative procedure from both a Bayesian and frequentist point of view.

In what follows we apply the ideas presented here to some standard problems usually handled by hypothesis testing methods.

2.1 Bernoulli Models

Let $\underset{\sim}{x}^{(N)}$ be a set of binary data bearing two distinct labels, i.e., $z_j = 1$ or 2, so that we incorporate this directly by writing $\underset{\sim}{x}^{(N)} = (\underset{\sim}{x}_1^{(N_1)}, \underset{\sim}{x}_2^{(N_2)})$. We then wish to assess the relevance of the label for generating the response. In the formal language of statistical theory we assume two alternative hypotheses (models) i.e., two populations such that under π_i, $P(X_{ij} = 1) = \theta_i = 1 - P(X_{ij} = 0)$ for $j = 1,\ldots,N_i$, $i = 1,2$ where X_{ij} are independently distributed. Hence under

$$M_1: \quad \pi_1 = \pi_2 = \pi \quad \text{or} \quad \theta_1 = \theta_2 = \theta$$

$$M_2: \quad \pi_1 \neq \pi_2 \quad \text{or} \quad \theta_1 \neq \theta_2 .$$

Under M_1 and a prior uniform density for $\theta \in [0,1]$ we can easily compute the predictive density for a future observation to be

$$P(X=1 \mid \underset{\sim}{x}^{(N)}, \pi) = \frac{r+1}{N+2} = 1 - P(X=0 \mid \underset{\sim}{x}^{(N)}, \pi) \tag{2.1.1}$$

where r is the number of ones in the set $\underset{\sim}{x}^{(N)}$. Under M_2 we compute, assuming θ_1 and θ_2 are a priori independent and uniformly distributed, the predictive probability of a future observation from π_i to be

$$P(X=1 \mid \underset{\sim}{x}^{(N)}, \pi_i) = \frac{r_i+1}{N_i+2} = 1 - P(X=0 \mid \underset{\sim}{x}^{(N)}, \pi_i) \tag{2.1.2}$$

where r_i is the number of ones in the set $\underset{\sim}{x}_i^{(N_i)}$.

Now we compute the product of the predicting sample reuse densities under the two alternative models and obtain

$$L_1 = \left(\frac{r}{N+1}\right)^{r} \left(\frac{N-r}{N+1}\right)^{N-r}$$

$$L_2 = \left(\frac{r_1}{N_1+1}\right)^{r_1} \left(\frac{N_1-r_1}{N_1+1}\right)^{N_1-r_1} \left(\frac{r_2}{N_2+1}\right)^{r_2} \left(\frac{N_2-r_2}{N_2+1}\right)^{N_2-r_2} \qquad (2.1.3)$$

These may be regarded as relative measures of the predictive plausibility of the two models due entirely to the data.

2.2 Exponential Models

Suppose under model M_1 we assume the data $\underset{\sim}{x}^{(N)}$ have been independently generated from a simple exponential distribution with density

$$f(x|\lambda) = \lambda e^{-\lambda x} \tag{2.2.1}$$

and under model M_2, assuming two distinct labels such that $x^{(N)} = (x_1^{(N_1)}, x_2^{(N_2)})$ where $x_i^{(N_i)} = (x_{i1},\ldots,x_{iN_i})$,

$$f(x|\lambda_i,\pi_i) = \lambda_i e^{-\lambda_i x} \qquad i = 1,2 . \tag{2.2.2}$$

Again taking a hint from a Bayesian analysis, we use vague priors of the type $g(\lambda_i) \propto \frac{1}{\lambda_i}$, and compute the predictive density of a future observation under

$$\begin{aligned} &M_1;\ \pi_1 = \pi_2 = \pi:\quad f(x|x^{(N)},\pi) = N(N\bar{x})^N/(N\bar{x} + x)^{N+1} \\ &M_2;\ \pi_1 \neq \pi_2:\quad f(x|x^{(N)},\pi_i) = N_i(N_i\bar{x}_i)^{N_i}/(N_i\bar{x}_i+x)^{N_i+1} \end{aligned} \tag{2.2.3}$$

where $\bar{x}_i = N_i^{-1}\sum_{i=1}^{N_i} x_i$ and $\bar{x} = N^{-1}(N_1\bar{x}_1 + N_2\bar{x}_2)$.

We shall use (2.2.3) as the basis for the sample reuse predicting densities (quasi-Bayes) though we note in passing that we could also use,

$$f(x|x^{(N)},\pi_i) = \hat{\lambda}_i e^{-\hat{\lambda}_i x} \tag{2.2.4}$$

where $\hat{\lambda}_i = \bar{x}_i^{-1}$ is the m.l.e. of λ. This was previously termed quasi-likelihood.

Hence applying the PSR criterion, we choose the larger of

$$L_1 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \frac{(N-1)[N\bar{x} - x_{ij}]^{N-1}}{(N\bar{x})^N} \quad \text{for } x_{ij} > 0 \qquad (2.2.5)$$

$$= 0 \quad \text{otherwise}$$

$$L_2 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \frac{(N_i-1)[N_i\bar{x}_i - x_{ij}]^{N_i-1}}{[N_i\bar{x}_i]^{N_i}} \quad \text{for } x_{ij} > 0 \qquad (2.2.6)$$

$$= 0 \quad \text{otherwise .}$$

If we had used (2.2.4) then

$$\hat{L}_1 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \left(\frac{N-1}{N\bar{x}-x_{ij}}\right) e^{-\frac{(N-1)x_{ij}}{N\bar{x}-x_{ij}}} \quad \text{for } x_{ij} > 0 \qquad (2.2.7)$$

$$= 0 \quad \text{otherwise}$$

$$\hat{L}_2 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \left(\frac{N_i-1}{N_i\bar{x}_i - x_{ij}}\right) e^{-\frac{(N_i-1)x_{ij}}{N_i\bar{x}_i-x_{ij}}} \quad \text{for } x_{ij} > 0 \qquad (2.2.8)$$

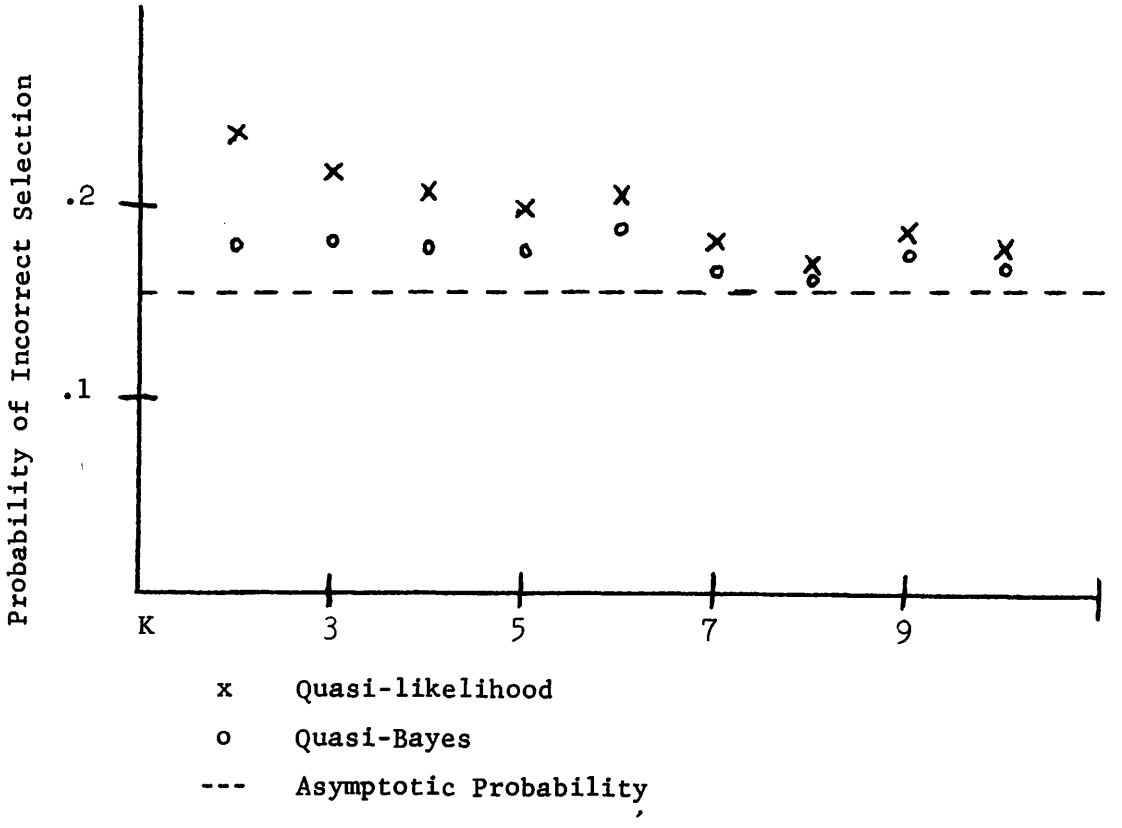
$$= 0 \quad \text{otherwise.}$$

There is some interest in comparing these two criteria: The quasi-Bayes (L_1 and L_2) and the quasi-likelihood ($\hat{L}_1$ and $\hat{L}_2$). It can be shown under M_1 that as the sample sizes increase that L_1 and $\hat{L}_1$ converge to a common value, as do L_2 and $\hat{L}_2$. Appealing to the aforementioned result of Stone (1976) the two different criteria are then asymptotically equivalent to Akaike's criterion and have asymptotic significance level $P\{\chi^2_1 > 2\}$ under M_1.

For small sample sizes a Monte Carlo experiment was performed to determine the rate at which the two criteria selected M_2 for data generated under M_1. For $N_1 = N_2 = K$, 10,000 samples each of size $K = 2(1)10$ were generated and the two criteria were computed for each sample. The nine two-by-two tables indicating the selection rates are given in Appendix I. For each sample size, Figure 2.2.1 indicates the probability of incorrect selection for each criterion together with the asymptotic value: $P\{\chi_1^2 > 2\} = 1.57$. It should be noted that in every case the error rate for the quasi-Bayes criterion was smaller than the rate for the quasi-likelihood criterion.

Figure 2.2.1

Comparison of Quasi-Bayes and Quasi-Likelihood Selection Criteria for the Exponential Model M_1 [1]



[1] Based on 10,000 samples

To illustrate the use of this quasi-Bayes criterion some examples are presented. First, consider the following data from Gross and Clark (1975, Table 7.1).

Table 2.2.1

Time to relief (in minutes) of headache pain

Patient	Standard Treatment	New Treatment
1	8.4	6.9
2	7.7	6.8
3	10.1	10.3
4	9.6	9.4
5	9.3	8.0
6	9.1	8.8
7	9.0	6.1
8	7.7	7.4
9	8.1	8.0
10	5.3	5.1

For illustrative purposes, they ignored the fact that these data were paired and assumed they were dealing with two independent samples. The F ratio $\bar{x}_{standard}/\bar{x}_{new}$, was computed to be 1.10. Since $P\{F_{20,20} > 1.10\} = .42$, a two-sided test has probability .84. Using the methods presented here, the following values were computed: $\log L_1 = -12.64$ and $\log L_2 = -13.57$. Thus both the quasi-Bayes criterion and the usual likelihood ratio test do not disagree that M_1 should be preferred.

The second set of data is taken from Davis (1952, Appendix I), and illustrates the fact that this approach is easily extended to several populations. It is counts of the number of correct bank statement entries between errors plus one listed in order of occurrence over a period of 10 days for five clerks. The first 26 observations on four selected clerks are presented in Table 2.2.2. Consider first a comparison of clerks #2 and #3. Log L_1 has the value -393.62 and log L_2 -394.45. So the PSR method prefers the model that these two clerks are the same.

Table 2.2.2

Number of Correct Ledger Entries Between Errors

1st Clerk	2nd Clerk	3rd Clerk	5th Clerk
734	451	726	149
121	3	883	74
404	1116	142	170
646	1143	196	2
1072	447	14	129
148	630	1905	3
312	37	456	65
773	2031	2565	44
43	1786	610	204
1102	659	1263	333
111	151	347	60
641	210	881	11
754	1426	1214	60
598	72	248	20
86	699	195	608
2138	426	548	19
150	1040	234	64
1047	277	1096	113
907	72	530	413
165	1286	338	75
166	235	356	22
6	625	217	403
94	493	195	299
1023	2	77	396
903	756	392	6
355	1460	3114	156

Now consider three clerks, #1, #2, and #3. With three possible populations there are a total of five possible partitions: (123), (1)(23), (2)(13), (3)(12), (1)(2)(3). Computing log L under each of the five partitions yields -584.80, -584.95, -585.40, -585.28, -585.79, respectively. Thus the model which says that all three clerks are the same is preferred. Another example considers clerks #2, #3, and #5. For the partitions (235), (2)(35), (3)(25), (5)(23), (2)(3)(5) the values of log L are -566.77, -565.93, -564.71, -551.00, -551.84, respectively. The model with #2 and #3 the same and #5 different is preferred.

The computer programs which computed the values of log L for the exponential model with two populations and with three populations are given in Appendices II and III respectively.

2.3 Normal Models

Here we shall assume three possible models but only two distinct labels so that the data could have arisen from two normal populations specified by π_i, $i = 1,2$ with density

$$f(x|\mu_i,\sigma_i^2,\pi_i) = (\sigma_i\sqrt{2\pi})^{-1}e^{-\frac{1}{2\sigma_i^2}(x-\mu_i)^2} \qquad (2.3.1)$$

but permitting the three following models;

M_1: $\pi_1 = \pi_2$ or $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2$

M_2: $\pi_1 \neq \pi_2$ or $\mu_1 \neq \mu_2$ but $\sigma_1^2 = \sigma_2^2$

M_3: $\pi_1 \neq \pi_2$ or $\mu_1 \neq \mu_2$ and $\sigma_1^2 \neq \sigma_2^2$.

Again taking our hint for the predicting densities from a Bayesian predictive analysis with the usual improper prior $g(\mu_i,\sigma_i) \propto \frac{1}{\sigma_i}$ we can compute the following predictive densities, c.f. Geisser (1964),

$$f(x|\underset{\sim}{x}^{(N)},M_1,\pi) \propto (1+\frac{N(x-\bar{x})^2}{(N^2-1)t^2})^{-\frac{N}{2}} \qquad (2.3.2)$$

$$f(x|\underset{\sim}{x}^{(N)},M_2,\pi_i) \propto (1+\frac{N_i(x-\bar{x}_i)^2}{(N_i+1)(N-2)s^2})^{-\frac{N-1}{2}} \qquad (2.3.3)$$

$$f(x|\underset{\sim}{x}^{(N)},M_3,\pi_i) \propto (1+\frac{N_i(x-\bar{x}_i)^2}{(N_i^2-1)s_i^2})^{-\frac{N_i}{2}} \qquad (2.3.4)$$

where

$$t^2 = (N-1)^{-1}\sum\sum_{ij} (x_{ij}-\bar{x})^2,\ s^2 = (N-2)^{-1}\sum_{i=1}^{2}(N_i-1)s_i^2,\ s_i^2 = (N_i-1)^{-1}\sum_{j=1}^{N_i}(x_{ij}-\bar{x}_i)^2 . \tag{2.3.5}$$

Using 2.3.2 - 2.3.4 as the means of forming predicting densities we then obtain

$$L_1 = \prod_{i=1}^{2}\prod_{j=1}^{N_i} \left(\frac{N-1}{\pi(N-2)N}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N-2}{2})t_{(ij)}} \left(1 + \frac{(N-1)(x_{ij}-\bar{x}_{(ij)})^2}{N(N-2)t^2_{(ij)}}\right)^{-\frac{N-1}{2}} \tag{2.3.6}$$

where

$$\bar{x}_{(ij)} = (N-1)^{-1}\sum_{k,t}^{(ij)} x_{kt};\ t^2_{(ij)} = (N_1+N_2-2)^{-1}\sum_{k,t}^{(ij)}(x_{kt}-\bar{x}_{(ij)})^2 \tag{2.3.7}$$

and $\sum^{(ij)}$ represents the sum over all values except x_{ij} ,

$$L_2 = \prod_{i=1}^{2}\prod_{j=1}^{N_i} \left(\frac{N_i-1}{\pi(N-3)N_i}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{N-2}{2})}{\Gamma(\frac{N-3}{2})s_{(i,j)}} \left[1 + \frac{(N_i-1)(x_{ij}-\bar{x}_{i(j)})^2}{N_i(N-3)s^2_{(ij)}}\right]^{-\frac{N-2}{2}} \tag{2.3.8}$$

where

$$\bar{x}_{i(j)} = (N_i-1)^{-1}\sum_{\substack{t=1\\ t\neq j}}^{N_i} x_{it};\ s^2_{(ij)} = (N_1+N_2-3)^{-1}[(N_i-2)s^2_{i(j)} + (N_{3-i}-1)s^2_{3-i}] ; \tag{2.3.9}$$

$$s^2_{i(j)} = (N_i-2)^{-1}\sum_{\substack{t=1\\ t\neq j}}^{N_i}(x_{it}-\bar{x}_{i(j)})^2;$$

and

$$L_3 = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \left(\frac{N_i-1}{\pi(N_i-2)N_i}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{N_i-1}{2}\right)}{\Gamma\left(\frac{N_i-2}{2}\right)s_{i(j)}} \left(1 + \frac{(N_i-1)(x_{ij}-\bar{x}_{i(j)})}{N_i(N_i-2)s^2_{i(j)}}\right) - \frac{N_i-1}{2} \; . \tag{2.3.10}$$

As an illustrative example here consider the following data from Lindley (1965, p. 124, Problem 11).

Table 2.3.1

Values of the Cephalic Index for Two Random Samples of Skulls

Sample I	74.1, 77.7, 74.4, 74.0, 73.8, 79.3, 75.8,
	82.8, 72.2, 75.2, 78.2, 77.1, 78.4, 76.3, 76.8
Sample II	70.8, 74.9, 74.2, 70.4, 69.2, 72.2, 76.8,
	72.4, 77.4, 78.1, 72.8, 74.3, 74.7

The values of log L were computed for these data under the three possible models yielding: $\log L_1 = -72.05$, $\log L_2 = -69.70$, and $\log L_3 = -70.95$. So the sample reuse criterion clearly opts for model M_2 where $\mu_1 \neq \mu_2$ but $\sigma_1^2 = \sigma_2^2$.

The usual hypothesis test comparing M_2 and M_3 leads to a value of 1.0479 for the F statistic with 13 and 15 degrees of freedom. The two-sided test does not reject model M_2. Conditional on this result the test comparing M_1 and M_2 leads to a value of 5.2798 for the t statistic with 26 degrees of freedom. The null hypothesis is easily rejected. In this example hypothesis testing and the sample reuse criterion lead to the same result: M_2.

The computer program which calculated the values of $\log L_1$, $\log L_2$, and $\log L_3$ is given in Appendix IV.

We further note that this model selection approach can be easily extended to multivariate normal populations. This may be accomplished by utilizing reused versions of the densities given by Geisser (1964, eqs. 3.21, and 3.33).

A common problem in multiple regression situations involves the "optimal" choice of some subset of potential independent variables. Consider the normal linear model; X_j independently distributed with mean $\underset{\sim}{z}_j'\underset{\sim}{\beta}$, $j = 1,\ldots,N$ and common variance σ^2 where $\underset{\sim}{z}_j' = (z_{1j},\ldots,z_{qj})$, $\underset{\sim}{\beta}' = (\beta_1,\ldots,\beta_q)$, $\underset{\sim}{Z} = (\underset{\sim}{z}_1,\ldots,\underset{\sim}{z}_N)$. Assuming the improper prior density $g(\beta,\sigma) \propto \frac{1}{\sigma}$ one can obtain the predictive density of a future observation, e.g., Geisser (1965). This quasi-Bayes reused predicting density is computed for some subset of arbitrary size k of the q regression coefficients, without loss of generality the first k . M_k then represents the model which includes only the first k regression coefficients.

$$L_k = \prod_{j=1}^{N} \left(\frac{c_j}{a^2_{(j)}\pi}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{N-k}{2})}{\Gamma(\frac{N-k-1}{2})} \left(1 + \frac{c_j(x_j - \underset{\sim}{z}_j'\underset{\sim}{b}_{(j)})^2}{a^2_{(j)}}\right)^{-\frac{N-k}{2}} \qquad (2.3.11)$$

$$\underset{\sim}{x}_{(j)}' = (x_1,\ldots,x_{j-1}, x_{j+1},\ldots,x_N)$$

$$\underset{\sim}{Z}_{(j)} = (\underset{\sim}{z}_1,\ldots,\underset{\sim}{z}_{j-1}, \underset{\sim}{z}_{j+1},\ldots,\underset{\sim}{z}_N)$$

$$\underset{\sim}{b}_{(j)} = (\underset{\sim}{Z}_{(j)}\underset{\sim}{Z}'_{(j)})^{-1}\underset{\sim}{Z}_{(j)}\underset{\sim}{x}_{(j)}$$

$$a^2_{(j)} = (\underset{\sim}{x}_{(j)} - \underset{\sim}{Z}'_{(j)}\underset{\sim}{b}_{(j)})'(\underset{\sim}{x}_{(j)} - \underset{\sim}{Z}'_{(j)}\underset{\sim}{b}_{(j)})$$

$$c_j = 1 - \underset{\sim}{z}_j'(\underset{\sim}{Z}\underset{\sim}{Z}')^{-1}\underset{\sim}{z}_j$$

where k replaces q .

By appropriate renumbering one could of course compare any subset S_1 of the q variables with any other, say S_2, by calculating $L(S_1)$ and $L(S_2)$ and preferring the model with the larger product.

For the normal multivariate regression case one can utilize reused quasi-Bayes predicting densities based on Geisser (1965, eq. 4.17).

To illustrate the use of this regression criteria consider the classical Hald data from Draper & Smith (1966, Appendix B) given in Table 2.3.2.

Table 2.3.2

Hald Data

X_0	X_1	X_2	X_3	X_4	Y
1	7	26	6	60	78.5
1	1	29	15	52	74.3
1	11	56	8	20	104.3
1	11	31	8	47	87.6
1	7	52	6	33	95.9
1	11	55	9	22	109.2
1	3	71	17	6	102.7
1	1	31	22	44	72.5
1	2	54	18	22	93.1
1	21	47	4	26	115.9
1	1	40	23	34	83.8
1	11	66	9	12	113.3
1	10	68	8	12	109.4

The values of the criterion $\log L_K$ were computed for the 16 possible regressions based on $X_1,\ldots,X_4$. They are reported together with the residual mean square for each regression from Draper and Smith in Table 2.3.3.

Table 2.3.3

Regression Variables	Residual Mean Square	log L_K
1	115.06	-65.86
2	82.39	-64.50
3	176.31	-68.75
4	80.35	-63.77
12	5.79	-46.23
13	122.71	-66.28
14	7.48	-48.13
23	41.54	-59.90
24	86.89	-64.67
34	17.57	-54.00
123	5.35	-45.57
124	5.33	-45.29
134	5.65	-45.71
234	8.20	-48.33
1234	5.98	-45.80

It should be noted that the ordering of the subsets of regression variables induced by log L_K is nearly identical to the ordering induced by the residual mean square. A computer program written in APL which computed the values of log L_K is given in Appendix VII.

2.4 Non-nested Models

Previously the models we treated were those usually associated with standard statistical hypothesis testing paradigms and as such were nested. The parameters under one model were unrestricted, while for other alternatives, the parameters were restricted in varying degrees to lower dimensional spaces. As we have mentioned earlier, full Bayesian analyses are possible but invariably include lumps of probability for the lower dimensional spaces to enable posterior odds ratios to differ from 0 and ∞. When the models represent different distributional families whose parameters bear no direct relation to one another the posterior odds ratio will depend on the joint predictive density of the observations under the various models. We noted that this predictive density depended entirely on the prior assumptions and indicated our reluctance to be so heavily dependent on them. Also if one were to utilize improper priors then the joint predictive density would be improper and hence of very limited value. So that even though the non-nested situation appears to be more similar to the full Bayesian treatment it still differs in that it substitutes the product of conditional predictive densities for the joint predictive density in order to give more weight to the data. In this respect it is a synthesis of frequentist and Bayesian notions seasoned by the data analytic spice of sample reuse.

As an example of a non-nested situation we could postulate that under M_1, the set of observations $\underset{\sim}{x}^{(n)}$ was generated by a simple exponential distribution while under M_2 by a normal distribution. Hence we would then compare L_1 of (2.2.5) with L_1 of (2.3.6) to assess which assumption for the data is more appropriate.

If one's intention is to predict a single point for a future observation by using the mean of the predictive distribution then this assessment is irrelevant since under each model the predictive mean is $\bar{x}$. However other values such as the mode or median will differ rather sharply. More to the point - if we are interested, as well we should be when assumptions permit, in predicting an interval for a future observation then the two models will certainly provide rather different solutions.

3. Low Structure Selection

By a low structure situation we mean that no assumptions about the likelihood are made for one or more of the alternative models available for selection. This lack of specificity compels us to abandon the predicting density criterion and return to the more primitive point prediction function $x(\underset{\sim}{x}^{(N)},\underset{\sim}{z}^{(N)}z,M_k) = f_k(\underset{\sim}{x}^{(N)})$. Additionally we introduce a discrepancy function D_k , a summary measure of deviations of observed values x_j from their predicted values $\hat{x}_j(\underset{\sim(j)}{x}^{(N-1)},\underset{\sim}{z}^{(N)},M_k) = f_k(\underset{\sim(j)}{x}^{(N-1)})$. Minimization of D_k with respect to k leads to the selection of the "best" model. A formal description of such a low structure procedure useful in certain very basic statistical paradigms was presented by Geisser (1974, 1975). In this section we apply this work to some of the previously discussed models and relate this to other methods.

3.1 One or Two Groups?

One of the classical hypothesis testing situations is basically whether a recognizable label is relevant for a response. Consider a set of data $\underset{\sim}{x}^{(N)}$ so labeled that $x^{(N)} = (x_1^{(N_1)}, x_2^{(N_2)})$, dispensing with $z^{(N)}$, and we are to determine whether the label is informative. There is then the possibility of two populations π_1 and π_2 that may differ (at least with respect to location) or that for all intents and purposes do not, i.e., $\pi_1 = \pi_2 = \pi$. We shall make a choice on the basis of whether the data are better predicted when considered as one population or two. If we are to predict a future observation from a population that a set of data represent then we would usually, in the absence of other information, use some central value such as the sample, mean, mode or median. Here we shall use the sample mean as a predictor for a future observation. Under M_1: $\pi_1 = \pi_2$ we shall use $\bar{x}$ and under M_2: $\pi_1 \neq \pi_2$ we shall use $\bar{x}_i$ $i = 1,2$. For convenience we shall use squared predictive error as our discrepancy measure with all notation as defined in section (2.3) and with one-at-a-time omissions, we compute, under M_1

$$D_1 = N^{-1} \sum_{i=1}^{2} \sum_{j=1}^{N_i} (x_{ij} - \bar{x}_{(ij)})^2 = \frac{N}{(N-1)^2} [(N_1-1)s_1^2 + (N_2-1)s_2^2 + \frac{N_1 N_2}{N}(\bar{x}_1 - \bar{x}_2)^2] \tag{3.1.1}$$

and compare this to

$$D_2 = N^{-1} \sum_{i=1}^{2} \sum_{j=1}^{N_i} (x_{ij} - \bar{x}_{i(j)})^2 = N^{-1} [\frac{N_1^2}{N_1-1} s_1^2 + \frac{N_2^2}{N_2-1} s_2^2] , \tag{3.1.2}$$

the discrepancy when prediction is made using only the data with the same label.

Hence if $D_1 \leq D_2$ choose M_1, otherwise choose M_2. For the special case $N_1 = N_2 = K$ we find that $D_1 > D_2$ is equivalent to

$$\frac{K(\bar{x}_1 - \bar{x}_2)^2}{2s^2} > \frac{4K-3}{2(K-1)} \tag{3.1.3}$$

Note that the l.h.s. of (3.1.3) has an $F_{1,2(K-1)}$ distribution if all of the observations were independent and identically distributed normal random variables. Hence for equal size normal samples an F-test with varying α level emerges (see first row of Table 3.1).

The comparison $D_1 > D_2$ of (3.1.1) and (3.1.2) tends to

$$\frac{(\bar{x}_1 - \bar{x}_2)^2}{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}} > 2 \tag{3.1.4}$$

for large N_1 and N_2. If all the random variables are i.i.d., $\pi_1 = \pi_2$, and the first two moments exist then the quantity on the left tends to a χ_1^2 random variable as N_1 and N_2 increase. When the data are binary, i.e., $x_i = 0$ or 1 then the criterion becomes

$$(\hat{p}_1 - \hat{p}_2)^2 > \frac{2NN_1 - N - N_1}{N(N_1-1)^2}\hat{p}_1(1-\hat{p}_1) + \frac{2NN_2 - N - N_2}{N(N_2-1)^2}\hat{p}_2(1-\hat{p}_2) \tag{3.1.5}$$

where $\hat{p}_i = r_i/N_i$. When $N_1 = N_2 = K$ say, (3.1.5) can be written in a more familiar form

$$\frac{K(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}_1(1-\hat{p}_1) + \hat{p}_2(1-\hat{p}_2)} > \frac{(4K-3)K}{2(K-1)^2} \quad . \tag{3.1.6}$$

Specifically if one assumed that these data were independent Bernoulli variables all emanating from the same population then clearly the l.h.s. of (3.1.6) tends to a χ^2_1 variable as $K \to \infty$. Hence asymptotically this would behave as a significance test with $P[\chi^2_1 > 2] = .157$, the α rejection level for the null hypothesis. It should be noted that the standard χ^2 statistic for this problem is slightly different from the l.h.s. of (3.1.6). It has a denominator which is $2\hat{p}(1-\hat{p})$ where $\hat{p} = (r_1 + r_2)/2K$.

As examples of the use of this low structure criterion consider the data introduced in Section 2. The data from Gross and Clark (1975) given in Table 2.2.2 yields $D_1 = 2.2919$ and $D_2 = 2.3801$. So the one population model M_1 is preferred as it was when the quasi-Bayes criterion was used.

The data from Davis (1952) given in Table 2.2.3 yields the values $D_1 = 46.0238$ and $D_2 = 47.8248$ for a comparison of clerks #2 and #3. This is in agreement with the quasi-Bayes criterion for the exponential model which also indicated M_1 should be preferred.

As a final example of the two group comparison consider the data from Lindley (1965) in Table 2.3.1. The value of D_1 is 9.2545 and the value of D_2 is 7.8743. While the low structure criterion is not directly comparable with the quasi-Bayes criterion for the normal model, the result is similar: two populations are preferred to one. The computer program which calculated the values of D_1 and D_2 is given in Appendix V.

In order to have some understanding of how the criterion works we make some simple assumptions and instructive calculations. Suppose $\pi_1 \neq \pi_2$ as the alternative to $\pi_1 = \pi_2$ signifies that $E(X|\pi_1) = \mu_1 \neq \mu_2 = E(X|\pi_2)$ but that $var(X|\pi_i) = \sigma^2$ for $i = 1,2$. Then we calculate

$$E(D_1|\pi_1 = \pi_2) = \frac{N}{N-1}\sigma^2 \tag{3.1.7}$$

$$E(D_2|\pi_1 = \pi_2) = \frac{\sigma^2}{N}\left[\frac{N_1^2}{N_1-1} + \frac{N_2^2}{N_2-1}\right] \tag{3.1.8}$$

and easily find that, for all $N_i > 1$,

$$E(D_1|\pi_1 = \pi_2) < E(D_2|\pi_1 = \pi_2) . \tag{3.1.9}$$

Actually the right hand side is minimized for each fixed N when $|N_1 - N_2| = 0$ or 1. In particular if $N_1 = N_2 = K$ then both sides of equation (3.1.9) are evaluated as

$$\sigma^2(1+\frac{1}{2K-1}) < \sigma^2(1+\frac{1}{K-1}) . \tag{3.1.10}$$

If the specified alternative is true then

$$E(D_1|\pi_1 \neq \pi_2) = \frac{N}{N-1}\sigma^2 + \frac{N_1N_2}{(N-1)^2}(\mu_1 - \mu_2)^2 \tag{3.1.11}$$

while $E(D_2|\pi_1 \neq \pi_2) = E(D_2|\pi_1 = \pi_2)$. For large N_1 and N_2 $D_2 > D_1$ if

$$(\mu_1 - \mu_2)^2 \geq \sigma^2(\frac{1}{N_1} + \frac{1}{N_2}) + O[(\frac{1}{N_1} + \frac{1}{N_2})^2] . \tag{3.1.12}$$

A simple exact expression when $N_1 = N_2 = K$ is

$$(\mu_1 - \mu_2)^2 \geq \frac{\sigma^2(2K-1)}{K(K-1)} \approx \frac{2\sigma^2}{K} . \tag{3.1.13}$$

Hence on the average one would select M_2, i.e., $\pi_1 \neq \pi_2$ when it is true, if the squared difference of the population means is larger than the variance of the difference between two sample means of size N_1 and N_2 respectively. Hence as a model selector the criterion is less than perfect because it concerns itself with the squared error of prediction which depends on the variance, the sample sizes and the difference between the means of the two populations. All of these factors are involved in prediction error. Further it follows, from a trivial computation, that it is possible to have smaller prediction error using the grand sample mean as opposed to the appropriate group mean even when the population means differ. The point of course being that optimal point prediction is not equivalent to optimal model selection. If we use the methodology strictly for the prediction of a single observation from π_i, $i = 1,2$ then we would compute

$$D_{1i} = N_i^{-1}\sum_{j=1}^{N_i}(x_{ij}-\bar{x}_{(ij)})^2 = \frac{N^2}{(N-1)^2}\left[\frac{N_i-1}{N_i}s_i^2+(\bar{x}-\bar{x}_i)^2\right] \qquad (3.1.14)$$

and

$$D_{2i} = N_i^{-1}\sum_{j=1}^{N_i}(x_{ij}-\bar{x}_{i(j)})^2 = \frac{N_i}{N_i-1}s_i^2 \quad . \qquad (3.1.15)$$

Hence $D_{2i} < D_{1i}$ for $i = 1, 2$ if

$$\frac{(\bar{x}_i-\bar{x}_{3-i})^2}{s^2(\frac{1}{N_i}+\frac{1}{N_{3-i}})} \geq \frac{2NN_i-N-N_i}{N(N_i-1)} \quad . \qquad (3.1.16)$$

Again under the assumption that $\underset{\sim}{X}^{(N)}$ is a set of i.i.d. normal random variables the l.h.s. is an F_{1,N_i-1} random variable. Hence it is clear from the prediction point of view that in one instance the criterion can choose $\bar{x}_i$ to predict a future observation from π_i and at the same time choose $\bar{x}$ to predict from π_{3-i}. In other words the criterion for prediction will be consistent with selection if for $i = 1$ and 2,

$$\frac{(\bar{x}_1-\bar{x}_2)^2}{s^2(\frac{1}{N_1}+\frac{1}{N_2})} \geq \frac{2NN_i-N-N_i}{N(N_i-1)} \tag{3.1.17}$$

or the inequality is reversed for both $i = 1$ and 2. We note that the original criterion (3.1.1) and (3.1.2) which was used for model selection is just the weighted average, i.e., for $u = 1, 2$

$$D_u = N^{-1}\sum_{i=1}^{2} N_i D_{ui} \tag{3.1.18}$$

which avoids the previously mentioned inconsistency. In a sense we have adapted our predictive criterion for selective purposes.

Table 3.1.1

$$P\{F_{r-1,r(k-1)} > 2 + \frac{r-1}{r(k-1)}\} = \alpha(r,k)$$

k		2	3	4	5	6	7	8	9	10	15	20	30	40	50	∞
r	2	.255	.208	.191	.183	.178	.175	.172	.170	.169	.165	.163	.161	.160	.159	.157
	3	.216	.178	.164	.157	.153	.150	.148	.146	.145	.142	.140	.138	.138	.137	.135
	4	.177	.146	.135	.129	.126	.124	.122	.121	.120	.117	.115	.114	.113	.113	.112
	5	.144	.119	.110	.106	.103	.101	.100	.099	.098	.096	.095	.094	.093	.093	.092
	6	.119	.098	.090	.087	.084	.083	.082	.081	.080	.079	.078	.077	.076	.076	.075
	7	.098	.080	.074	.071	.069	.068	.067	.067	.066	.065	.064	.063	.063	.063	.062
	8	.081	.066	.061	.059	.057	.056	.056	.055	.055	.053	.053	.052	.052	.052	.051
	9	.067	.055	.051	.049	.047	.047	.046	.046	.045	.044	.044	.043	.043	.043	.042
	10	.056	.046	.042	.040	.039	.039	.038	.038	.037	.037	.036	.036	.036	.036	.035
	15	.024	.019	.017	.016	.016	.016	.015	.015	.015	.015	.015	.015	.014	.014	.012
	20	.0102	.0079	.0072	.0068	.0066	.0065	.0064	.0064	.0063	.0062	.0061	.0061	.060	.060	.0059
	30	.0020	.0015	.0013	.0013	.0012	.0012	.0012	.0012	.0012	.0012	.0012	.0012	.0012	.0011	.0011

3.2 Many Groups and Regression

We shall here present the case of three groups and it shall serve as the paradigm for r groups. If we have N observations and three recognizable labels then we can list the 5 possible models and compute their average sample reuse discrepancies. The models are

M_1: $(\pi_1 = \pi_2 = \pi_3)$

M_{2u}: $(\pi_u \neq \pi_v = \pi_w)$ $u = 1,2,3$ and (v,w) is the set of remaining integers $\neq u$ with the convention $v < w$.

M_3: $(\pi_1 \neq \pi_2 \neq \pi_3)$,

with associated sample reuse discrepancies

$$D_1 = \frac{N}{(N-1)^2}\left[(N-3)s^2 + \sum_1^3 N_i(\bar{x}_i - \bar{x})^2\right] \tag{3.2.1}$$

where $N = \sum_{i=1}^{3} N_i$, $\bar{x} = N^{-1}\sum_{i=1}^{3} N_i\bar{x}_i$ and $(N-3)s^2 = \sum_{i=1}^{3}\sum_{j=1}^{N_i}(\bar{x}_{ij} - \bar{x}_i)^2$, and

$$D_{2u} = N^{-1}\left[\frac{(N_v+N_w)^2}{(N_v+N_w-1)^2}\left\{(N_v-1)s_v^2+(N_w-1)s_w^2+\frac{N_vN_w}{N_v+N_w}(\bar{x}_v-\bar{x}_w)^2\right\}+\frac{N_u^2}{(N_u-1)}s_u^2\right] \tag{3.2.2}$$

for $u = 1,2,3$ and (v,w) defined as above,

$$D_3 = N^{-1}\sum_{i=1}^{3}\frac{N_i^2}{(N_i-1)}s_i^2 .$$

The smallest then of D_1, D_{21}, D_{22}, D_{23}, D_3 indicates the appropriate model or grouping to be chosen.

To illustrate the use of this criterion consider the data in Table 2.2.3. As in the exponential case, with the three clerks #1, #2, #3 there are five possible partitions (123), (1)(23), (2)(13), (3)(12), (1)(2)(3). The values of D_1, D_{21}, D_{22}, D_{23}, D_3 are respectively 386929, 390738, 396169, 397393, 402744. The smallest is D_1. Thus the model which says all three clerks are the same is preferred (in agreement with the quasi-Bayes criterion of Section 2.2). Consider now the three clerks #2, #3, and #5. The values of D are 379824, 376061, 370716, 315863, 327867 respectively. Thus the model which says that clerks #2 and #3 are the same and #5 is different is preferred as it was in Section 2.2.

To throw some light on this criterion we perform some further computations. For the case of equal sample size $N_i = K$, the formulas simplify to

$$D_1 = \frac{3K}{(3K-1)^2}\left[3(K-1)s^2 + K\sum_i (\bar{x}_i-\bar{x})^2\right]$$

$$D_{2u} = (3K)^{-1}\left[\left(\frac{2K}{2K-1}\right)^2\left\{(K-1)(s_v^2+s_w^2)+\frac{K}{2}(\bar{x}_v-\bar{x}_w)^2\right\} + \frac{K^2}{(K-1)}\,s_u^2\right] \quad (3.2.3)$$

$$D_3 = \frac{K}{K-1}\,s^2$$

Some comparisons can be made in terms of familiar statistics thus $D_1 > D_3$ if

$$\frac{K\sum_i(\bar{x}_i-\bar{x})^2}{2s^2} > \frac{6K-4}{3K-3} \quad (3.2.4)$$

and $D_{2u} > D_3$ if

$$\frac{K(\bar{x}_v - \bar{x}_w)^2}{(s_v^2 + s_w^2)} > \frac{4K-3}{2K-2} \quad . \tag{3.2.5}$$

However the comparison of D_1 with D_{2u} does not lend itself to a familiar statistic though for large K, $D_1 > D_{2u}$, approximately, if

$$\frac{K(2\bar{x}_u - \bar{x}_v - \bar{x}_w)^2}{(4s_u^2 + s_v^2 + s_w^2)} \geq 2 \quad . \tag{3.2.6}$$

If only M_1 and M_3 were tenable then the comparison of D_1 and D_3 under the usual normal assumptions when M_1 holds is again an F test since $D_1 > D_3$ is equivalent to

$$F_{2,3(K-1)} = \frac{K \sum_{i=1}^{3} (\bar{x}_i - \bar{x})^2}{2s^2} > \frac{6K-4}{3K-3} \quad . \tag{3.2.7}$$

For r groups of equal size this can be easily extended for comparing the two alternatives, all groups the same as opposed to all groups distinct, since the method asserts that the model, all groups the same is to be rejected whenever

$$F_{r-1,r(K-1)} = \frac{K \sum_{i=1}^{r} (\bar{x}_i - \bar{x})^2}{\frac{1}{(r-1)^2}} > 2 + \frac{r-1}{r(K-1)} \quad . \tag{3.2.8}$$

As $K \to \infty$ this tends to $(r-1)^{-1}\chi^2_{r-1}$ under the usual null hypothesis and other fairly general conditions so that it is equivalent to a significance level

$$\alpha(r) = P[\chi^2_{r-1} > 2(r-1)]. \tag{3.2.9}$$

From the Table 3.1.1 it is clear that $\alpha(r)$ tends to 0 monotonically in r. Again if the customary normal theory associated with the analysis of

variance holds the probability $\alpha(r,K)$ say, that the l.h.s. of (3.2.8) exceeds the r.h.s. under the null hypothesis is obtained from the appropriate F distribution. The values for $\alpha(r,K)$ are given in Table 3.1.1. The table indicates that $\alpha(r,K)$ is a monotonically decreasing function of r and K. This implies that if in fact the groups are all the same it becomes increasingly more difficult to assert that they are all distinct.

For the general case where N_i are arbitrary and again the only tenable models are M_1: $\pi_1 = \pi_2 = \ldots = \pi_r$ versus M_2: $\pi_1 \neq \ldots \neq \pi_r$ then

$$D_1 = \frac{N}{(N-1)^2}\left[(N-r)s^2 + \sum_{i=1}^{r} N_i(\bar{x}_i - \bar{\bar{x}})^2\right]$$

$$D_2 = N^{-1}\sum_{i=1}^{r} \frac{N_1^2}{(N_i-1)}\, s_i^2 \quad . \tag{3.2.10}$$

Of course one could enumerate all the intermediate models and compute their predictive discrepancy for a given r. In computing the discrepancy for any intermediate model there is basically only one algorithm necessary. Suppose r is assumed to be partitioned into m subgroups $r = \sum_{t=1}^{m} r_t$ where $r_t \geq 1$ then each such subgroup contributes to the total discrepancy

$$D(r_t) = \frac{N^2(r_t)}{(N(r_t)-1)^2}\left\{\left[N(r_t)-r_t\right]s^2(r_t) + \sum_{u=1}^{r_t} N_{i_u}(\bar{x}_{i_u} - \bar{x}(r_t))^2\right\} \tag{3.2.11}$$

where $\sum_{u=1}^{r_t} N_{i_u} = N(r_t)$, $\bar{x}_{i_u} = N_{i_u}^{-1}\sum_{k=1}^{N_{i_u}} x_{ki_u}$, $\bar{x}(r_t) = N^{-1}(r_t)\sum_{u=1}^{r_t} N_{i_u}\bar{x}_{i_u}$, and

$(N(r_t)-r_t)s^2(r_t) = \sum_{u=1}^{r_t}\sum_{k=1}^{N_{i_u}}(x_{ki_u} - \bar{x}_{i_u})^2$. Note if $r_s = 1$ then

$$D(r_s) = \frac{N^2(r_s)}{N(r_s)-1}\, s^2(r_s) \tag{3.2.12}$$

where here r_s refers to a single subscript from among the set $(1,2,\ldots,r)$.

Hence for this partition, say the P-th of r into m subgroups, the total discrepancy is

$$D_{Pp} = N^{-1} \sum_{i=1}^{r} D_{Pp}(r_t) \ , \tag{3.2.13}$$

where the second index on D refers to the particular permutation, p, within the P-th partition as each permutation leads to its own discrepancy. The total number of possibilities, partitions and permutations within partitions, increases so rapidly with r as to preclude their full enumeration and computation. Of course an investigator can choose some small subset which he believes are tenable for his purposes and compare their average discrepancies and thus make his selection on this limited basis. The usual analysis of variance selects out one partition and purports to be an omnibus test against all others.

For the comparison of various subsets of the q independent variables in the multiple regression case discussed in Section 2.3 we can compute for the first k (by reordering) an average discrepancy

$$D_k = N^{-1} \sum_{j=1}^{N} (x_j - \underset{\sim}{z}_j' \underset{\sim}{b}_{(j)})^2 \ . \tag{3.2.14}$$

This particular discrepancy measure was suggested by Allen (1971) and termed PRESS.

If some modest moment assumptions are made a weighted discrepancy

$$W_k = (N-k-3) \sum_{j=1}^{N} \frac{(x_j - \underset{\sim}{z}_j' \underset{\sim}{b}_{(j)})^2 c_j}{a^2_{(j)}} \tag{3.2.15}$$

may be more useful. In either case our choice for a particular set of independent variables depends on the smallest discrepancy amongst those compared.

Using the Hald data again as an example the criterion D_k was computed for the 15 possible regressions. The results are given in Table 3.2.1.

Table 3.2.1

Selection Criterion for Regression

Regression Variables	D_k
1	130.74
2	92.47
3	201.26
4	91.86
12	7.22
13	170.62
14	9.32
23	53.98
24	112.45
34	22.62
123	6.92
124	6.57
134	7.27
234	11.30
1234	8.49

Notice that again, as in the high structure situation, the ordering of the subsets of variables induced by the criterion D_k is very similar to that induced by the residual mean square. The APL program which computed the values of the criteria is given in Appendix VII.

3.3 Remarks on the Use of Predictive Function and Discrepancy

Throughout this section we have used the mean and squared error for predictive function and discrepancy respectively. These certainly have a long tradition in statistics as being sensible for a wide variety of cases especially when distributional assumptions are unspecifiable. Our only bias for these functions of the observations is that they do lead to fairly convenient algorithms. Now it may be that for prediction, some other functions are better but we must bear in mind that our goal here is not strictly prediction but selection. In view of this we would like to present a simple case which demonstrates how one can use a criterion which, though it may predict very well, does rather poorly from the point of view of selection.

Consider predicting a future binary random variable X with known $\theta = Pr(X=1) = 1-Pr(X=0)$. Suppose we use as a criterion, the maximization of the proportion of correct predictions. Then by considering the expected proportion of correct guesses it is clear that to guess anything but 0 or 1 would be inadmissible since it cannot add anything to the numerator of this fraction. Hence in its most general setup we need merely predict $X = 1$ with probability q and $X = 0$ with probability $1-q$. Therefore

$$\max_{q} E[\text{proportion of correct predictions} \mid \theta] = \max_{q}[q(2\theta-1) + 1-\theta] = \max[\theta, 1-\theta] \tag{3.3.1}$$

with solution $q = 1$ if $\theta \geq 1/2$, and $q = 0$ otherwise. Hence the optimal predictor is the mode or median. However one may use another criterion namely; minimization of squared predictive error

$$\min_{a} E(X-a)^2 = \theta(1-\theta) \qquad (3.3.2)$$

with solution $a = \theta$ so that here the optimal predictor is the mean. These are simple and well known facts but we now relate them to the problem of sample reuse model selection. Suppose a predictive criterion is invoked to ascertain whether two sets of data, provisionally distinguished only by a label, are in fact best treated as one or two populations. This may actually depend on whether we are interested in some underlying structure or more directly in predicting future observations from one or both labeled sets. One way of assessing this is to use an appropriate predictor on the data set itself with some criterion of comparative prediction when the data set is treated as one or two populations. In order to predict an observation we calculate the appropriate predictor without the observation. Then,as usual, a discrepancy of predicted from observed for each observation is calculated. These discrepancies are combined for all observations and compared when executed in both one or two population models. In order to make the point expeditiously, assume equal and even samples each of size $2J$. Suppose then we have r_i 1's and $2J-r_i$ 0's $i = 1,2$. We shall use median or mode as predictor, i.e., 0 or 1 and assume that the sample values r_1 and r_2 are both $< J-1$ or both $> J+1$ to avoid trivial complications. It is clear that adding the total number of correct predictions from the samples handled individually, i.e., assuming different populations, is exactly equivalent to the number of correct predictions when the populations are combined. Hence the predictor and predictive criterion--selecting the model which maximizes the number of correct predictions is insensitive to large regions of the possible values of r_1 and r_2 no matter how large J or how differ-

ent r_1 and r_2 are in these regions. While all this may be quite reasonable from a prediction point of view it is not sensible from the standpoint of model selection. Hence we would regard this combination of predictor and criterion as ineffective for model selection. On the other hand using the mean as predictor (which is guaranteed to predict incorrectly in all future cases, for $2 \leq r_i \leq 2J-2$, though on the average it is closest via squared error) can be eminently sensible in the model selection paradigm as indicated in Section 3.1.

References

Aitchison, J. (1975). Goodness of prediction fit. Biometrika, 62, 3, pp. 547-554.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Proceedings of the 2nd International Symposium on Information Theory, (B. N. Petrov and F. Czaki, eds., Akademiai Kiado, Budapest), pp. 267-281.

Allen, D. M. (1971). The prediction sum of squares as a criterion for selecting prediction variables. Technical Report No. 23, University of Kentucky.

Davis, D. J. (1952). An analysis of some failure data. J. Amer. Statist. Assoc., 47, pp. 113-150.

Draper, N. R. and Smith, H. (1966). Applied Regression Analysis, John Wiley and Sons, New York.

Geisser, S. (1964). Posterior odds for multivariate normal classification. Jour. of the Royal Stat. Soc., Part 1, Series B, pp. 69-76.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. Ann. Math. Stat., 36, pp. 150-159.

Geisser, S. (1966). Predictive discrimination. Multivariate Analysis, P. Krishnaiah, ed., Academic Press, pp. 149-163.

Geisser, S. (1971). The inferential use of predictive distributions. Foundations of Statistical Inference, V. Godambe and D. Sprott, Holt, Rinehart and Winston, pp. 456-469.

Geisser, S. (1974). A predictive approach to the random effect model. Biometrika, pp. 101-107.

Geisser, S. (1975). The predictive sample reuse method with applications. Jour. of the Amer. Stat. Assoc., Vol. 70, No. 350, pp. 320-328.

Gross, A. J. and Clark, V. A. (1975). Survival Distributions: Reliability Applications in the Biomedical Sciences, John Wiley & Sons, New York.

Kullback. S. and Leibler, R. A. (1951). On information and sufficiency. Ann. Math. Stat., 22, pp. 525-540.

Lee, J. C., and Geisser, S. (1972). Applications of growth curve prediction. University of Minnesota Technical Report No. 180.

Lindley, D. V. (1965). An Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2: Inference, Cambridge University Press, London.

Stone, M. (1976). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. (Unpublished paper).

Appendix I

Percent of Model Selections for the Quasi-Bayes and Quasi-Likelihood Criteria [1]

		Quasi-Bayes M_1	Quasi-Bayes M_2	Totals
Quasi-Likelihood k = 2	M_1	72.67	3.74	76.41
	M_2	9.39	14.20	23.59
	Totals	82.06	17.94	100.00

		Quasi-Bayes M_1	Quasi-Bayes M_2	Totals
Quasi-Likelihood k = 3	M_1	75.27	2.97	78.24
	M_2	6.63	15.13	21.76
	Totals	81.90	18.10	100.00

		Quasi-Bayes M_1	Quasi-Bayes M_2	Totals
Quasi-Likelihood k = 4	M_1	77.03	2.34	79.37
	M_2	5.09	15.54	20.63
	Totals	82.12	17.88	100.00

		Quasi-Bayes M_1	Quasi-Bayes M_2	Totals
Quasi-Likelihood k = 5	M_1	78.13	1.99	80.12
	M_2	4.22	15.66	19.88
	Totals	82.35	17.65	100.00

		Quasi-Bayes M_1	Quasi-Bayes M_2	Totals
Quasi-Likelihood k = 6	M_1	77.39	2.14	79.53
	M_2	3.76	16.71	20.47
	Totals	81.15	18.85	100.00

		Quasi-Bayes		
		M_1	M_2	Totals
Quasi-Likelihood k = 7	M_1	80.30	1.65	81.95
	M_2	3.09	14.96	18.05
	Totals	83.39	16.61	100.00

		Quasi-Bayes		
		M_1	M_2	Totals
Quasi-Likelihood k = 8	M_1	81.08	1.13	83.01
	M_2	2.66	14.33	16.99
	Totals	83.74	16.26	100.00

		Quasi-Bayes		
		M_1	M_2	Totals
Quasi-Likelihood k = 9	M_1	79.96	1.61	81.57
	M_2	2.68	15.75	18.43
	Totals	82.64	17.36	100.00

		Quasi-Bayes		
		M_1	M_2	Totals
Quasi-Likelihood k = 10	M_1	80.68	1.55	82.23
	M_2	2.61	15.16	17.77
	Totals	83.29	16.71	100.00

[1]Percentages based on 10,000 samples generated from the exponential model M_1 .

Appendix II

```
C THIS PROGRAM COMPUTES THE ONE GROUP AND TWO GROUP LOG
C PREDICTING DENSITIES BY THE SAMPLE REUSE METHOD FOR THE
C EXPONENTIAL MODEL
      DIMENSION X(30,2),N(2),XB(2)
      NG=2
C NUMBER OF GROUPS IS TWO
      DO 555 I=1,NG
555   N(I)=26
C NUMBER IN EACH GROUP IS 26
      DO 3 I=1,NG
C FOR EACH GROUP READ IN DATA AND COMPUTE GROUP SUM
      NI=N(I)
      READ(5,1)(X(J,I),J=1,NI)
1     FORMAT(10F5.0)
      XB(I)=0.
      DO 2 J=1,NI
2     XB(I)=XB(I)+X(J,I)
3     CONTINUE
C COMPUTE ONE GROUP LOG PREDICTING DENSITY
      XBN=0.
      NN=0.
      DO 4 I=1,NG
      NN=NN+N(I)
4     XBN=XBN+XB(I)
      SUMA=0.
      DO 6 I=1,NG
      NI=N(I)
      DO 5 J=1,NI
5     SUMA=SUMA+ALOG(XBN-X(J,I))
6     CONTINUE
      SUMA=(NN-1)*SUMA+NN*(ALOG(FLOAT(NN-1))-NN*ALOG(XBN))
C COMPUTE TWO GROUP LOG PREDICTING DENSITY
      SUMB=0.
      DO 8 I=1,NG
      SUMI=0.
      NI=N(I)
      XBI=XB(I)
      DO 7 J=1,NI
7     SUMI=SUMI+ALOG(XBI-X(J,I))
8     SUMB=(NI-1)*SUMI+NI*(ALOG(FLOAT(NI-1))-NI*ALOG(XBI))+SUMB
C PRINT LOG PREDICTING DENSITIES
      WRITE(6,9)SUMA,SUMB
9     FORMAT(2F12.0)
      STOP
      END
```

Appendix III

```
C THIS PROGRAM COMPUTES THE SAMPLE REUSE LOG PREDICTING
C DENSITIES FOR THREE GROUPS UNDER THE EXPONENTIAL MODEL.
C THERE ARE FIVE POSSIBLE PARTITIONS:  ONE SINGLE GROUP,
C THREE DISTINCT GROUPS, AND THREE DIFFERENT TWO GROUP
C PARTITIONS.
      DIMENSION X(30,3),N(3),XB(3),SUMB(3)
      NG=3
C NUMBER OF GROUPS IS THREE
      DO 555 I=1,NG
555   N(I)=26
C NUMBER IN EACH GROUP IS 26
      DO 3 I=1,NG
C FOR EACH GROUP READ IN DATA AND COMPUTE GROUP SUM
      NI=N(I)
      READ(5,1)(X(J,I),J=1,NI)
1     FORMAT(10F5.0)
      XB(I)=0.
      DO 2 J=1,NI
2     XB(I)=XB(I)+X(J,I)
3     CONTINUE
C COMPUTE ONE GROUP LOG PREDICTING DENSITY
      XBN=0.
      NN=0.
      DO 4 I=1,NG
      NN=NN+N(I)
4     XBN=XBN+XB(I)
      SUMA=0.
      DO 6 I=1,NG
      NI=N(I)
      DO 5 J=1,NI
5     SUMA=SUMA+ALOG(XBN-X(J,I))
6     CONTINUE
      SUMA=(NN-1)*SUMA+NN*(ALOG(FLOAT(NN-1))-NN*ALOG(XBN))
C COMPUTE LOG PREDICTING DENSITY FOR THREE DISTINCT GROUPS
      SUMC=0.
      DO 8 I=1,NG
      SUMI=0.
      NI=N(I)
      XBI=XB(I)
      DO 7 J=1,NI
7     SUMI=SUMI+ALOG(XBI-X(J,I))
8     SUMC=(NI-1)*SUMI+NI*(ALOG(FLOAT(NI-1))-NI*ALOG(XBI))+SUMC
C COMPUTE THE THREE DIFFERENT LOG PREDICTING DENSITIES
C CORRESPONDING TO THE THREE DIFFERENT TWO GROUP PARTITIONS
      DO 13 I=1,NG
      SUMB(I)=0.
      SUMI=0.
      NI=N(I)
      XBI=XB(I)
      DO 9 J=1,NI
```

```
9       SUMI=SUMI+ALOG(XBI-X(J,I))
        SUMB(I)=SUMB(I)+(NI-1)*SUMI+NI*(ALOG(FLOAT(NI-1))-
       1-NI*ALOG(XBI))
        SUMI=0.
        NII=0
        XBI=0.
        DO 10 II=1,NG
        IF(II.EQ.I)GOTO 10
        NII=NII+N(II)
        XBI=XBI+XB(II)
10      CONTINUE
        DO 12 II=1,NG
        IF(II.EQ.I)GOTO 12
        NI=N(II)
        DO 11 J=1,NI
11      SUMI=SUMI+ALOG(XBI-X(J,II))
12      CONTINUE
        SUMB(I)=(NII-1)*SUMI+NII*(ALOG(FLOAT(NII-1))-NII-
       1*ALOG(XBI))+SUMB(I)
13      CONTINUE
C PRINT LOG PREDICTING DENSITIES FOR FIVE POSSIBLE PARTITIONS
        WRITE(6,14)SUMA,SUMB,SUMC
14      FORMAT(5F12.6)
        STOP
        END
```

Appendix IV

```
C THIS PROGRAM COMPUTES THE THREE SAMPLE REUSE LOG PREDICTING
C DENSITIES FOR TWO GROUPS UNDER THE NORMAL MODEL:  1) EQUAL
C MEANS AND EQUAL VARIANCES, 2)UNEQUAL MEANS BUT EQUAL
C VARIANCES AND 3)UNEQUAL MEANS AND UNEQUAL VARIANCES.
      DIMENSION X(15,2),N(2),XB(2),XL(3),SS(2)
      NG=2
C NUMBER OF GROUPS IS TWO
      N(1)=15
C FIFTEEN OBSERVATIONS IN THE FIRST GROUP
      N(2)=13
C THIRTEEN IN THE SECOND
      NN=N(1)+N(2)
      NNM1=NN-1
      NNM2=NN-2
      NNM3=NN-3
      DO 3 I=1,NG
C READ IN THE DATA AND COMPUTE THE GROUP SUMS
      NI=N(I)
      READ(5,1)(X(J,I),J=1,NI)
1     FORMAT(10F5.1)
      XB(I)=0.
      DO 2 J=1,NI
2     XB(I)=XB(I)+X(J,I)
3     CONTINUE
      PI=2.0*ARCOS(0.)
      ALG1=ALGAMA(.5*FLOAT(NNM1))
      ALG2=ALGAMA(.5*FLOAT(NNM2))
      ALG3=ALGAMA(.5*FLOAT(NNM3))
      DO 4 I=1,3
4     XL(I)=0.
C COMPUTE ONE GROUP LOG PREDICTING DENSITY
      XBB=0.
      DO 5 I=1,NG
5     XBB=XBB+XB(I)
      DO 9 I=1,NG
      NI=N(I)
      DO 8 J=1,NI
      XBIJ=(XBB-X(J,I))/FLOAT(NNM1)
      S=(X(J,I)-XBIJ)**2
      T=-S
      DO 7 II=1,NG
      NII=N(II)
      DO 6 JJ=1,NII
6     T=T+(X(JJ,II)-XBIJ)**2
7     CONTINUE
      T=T/NNM2
8     XL(1)=XL(1)+.5*ALOG(FLOAT(NNM1)/(PI*FLOAT(NNM2*NN)))+ALG1-
     1-ALG2-.5*ALOG(T)-FLOAT(NNM1)*.5*ALOG(1.+(FLOAT(NNM1)*S-
     2/(FLOAT(NN*NNM2)*T)))
9     CONTINUE
```

```
C COMPUTE BOTH TWO GROUP LOG PREDICTING DENSITIES
      DO 11 I=1,NG
      NI=N(I)
      SS(I)=0.
      XBIN=XB(I)/NI
      DO 10 J=1,NI
10    SS(I)=SS(I)+(X(J,I)-XBIN)**2
11    CONTINUE
      DO 15 I=1,NG
      NI=N(I)
      NIM2=NI-2
      NIM1=NI-1
      XBI=XB(I)
      DO 14 J=1,NI
      XBIJ=(XBI-X(J,I))/NIM1
      S=(X(J,I)-XBIJ)**2
      T=-S
      DO 13 JJ=1,NI
13    T=T+(X(JJ,I)-XBIJ)**2
      S2IPJP=T/NIM2
      III=3-I
      S2IPJP=(T+SS(III))/NNM3
      XL(2)=XL(2)+.5*ALOG(FLOAT(NIM1)/(PI*FLOAT(NNM3*NI)))+ALG2-
     1-ALG3-.5*ALOG(S2IPJP)-.5*FLOAT(NNM2)*ALOG(1.+FLOAT(NIM1)*S-
     2/(FLOAT(NI*NNM3)*S2IPJP))
14    XL(3)=XL(3)+.5*ALOG(FLOAT(NIM1)/(PI*FLOAT(NIM2*NI)))-
     1+ALGAMA(.5*FLOAT(NIM1))-ALGAMA(.5*FLOAT(NIM2))-.5-
     2*ALOG(S2IPJP)-.5*FLOAT(NIM1)*ALOG(1.+FLOAT(NIM1)*S-
     3/(FLOAT(NI*NIM2)*S2IPJP))
15    CONTINUE
      WRITE(6,16)XL
16    FORMAT(3F12.6)
      STOP
      END
```

Appendix V

```
C THIS PROGRAM COMPUTES THE LOW STRUCTURE SAMPLE REUSE CRITERION
C FOR MODEL SELECTION WITH THE SAMPLE MEAN AS A PREDICTOR AND
C SQUARED ERROR AS A DISCREPANCY MEASURE FOR TWO GROUPS
      DIMENSION X(30,2),XB(2),N(2)
      NG=2
C NUMBER OF GROUPS IS TWO
      N(1)=15
C NUMBER IN FIRST GROUP IS 15
      N(2)=13
C NUMBER IN SECOND IS 13
      DO 3 I=1,NG
C FOR EACH GROUP READ IN THE DATA AND COMPUTE THE GROUP SUM
      NI=N(I)
      READ(5,1)(X(J,I),J=1,NI)
1     FORMAT(10F5.2)
      XB(I)=0.
      DO 2 J=1,NI
2     XB(I)=XB(I)+X(J,I)
3     CONTINUE
C COMPUTE THE ONE GROUP DISCREPANCY
      D1=0.
      XBN=0.
      NN=0
      DO 4 I=1,NG
      XBN=XBN+XB(I)
4     NN=NN+N(I)
      NNM1=NN-1
      DO 6 I=1,NG
      NI=N(I)
      DO 5 J=1,NI
5     D1=D1+(FLOAT(NN)*X(J,I)-XBN)**2
6     CONTINUE
      D1=D1/FLOAT(NN*NNM1**2)
C COMPUTE THE TWO GROUP DISCREPANCY
      D2=0.
      DO 8 I=1,NG
      FNI=FLOAT(N(I))
      FNIM1=FNI-1.
      XBI=XB(I)
      DO 7 J=1,NI
7     D2=D2+((FNI*X(J,I)-XBI)/FNIM1)**2
8     CONTINUE
      D2=D2/FLOAT(NN)
C PRINT THE DISCREPANCIES
      WRITE(6,9)D1,D2
9     FORMAT(2F12.6)
      STOP
      END
```

APPENDIX VI

```
C THIS PROGRAM COMPUTES THE LOW STRUCTURE SAMPLE REUSE CRITERION
C FOR MODEL SELECTION WITH THE SAMPLE MEAN AS A PREDICTOR AND
C SQUARED ERROR AS A DISCREPANCY MEASURE FOR THREE GROUPS
      DIMENSION X(30,3),XB(3),N(3),D2(3)
      NG=3
C NUMBER OF GROUPS IS THREE
      N(1)=26
C NUMBER IN FIRST GROUP IS 26
      N(2)=26
C NUMBER IN SECOND GROUP IS 26
      N(3)=26
C NUMBER IN THIRD IS 26
      DO 3 I=1,NG
C FOR EACH GROUP READ IN THE DATA AND COMPUTE THE GROUP SUM
      NI=N(I)
      READ(5,1)(X(J,I),J=1,NI)
1     FORMAT(10F5.0)
      XB(I)=0.
      DO 2 J=1,NI
2     XB(I)=XB(I)+X(J,I)
3     CONTINUE
C COMPUTE THE ONE GROUP DISCREPANCY
      D1=0.
      XBN=0.
      NN=0
      DO 4 I=1,NG
      XBN=XBN+XB(I)
4     NN=NN+N(I)
      NNM1=NN-1
      DO 6 I=1,NG
      NI=N(I)
      DO 5 J=1,NI
5     D1=D1+(FLOAT(NN)*X(J,I)-XBN)**2
6     CONTINUE
      D1=D1/FLOAT(NN*NNM1**2)
C COMPUTE THE THREE DIFFERENT TWO GROUP DISCREPANCIES
      DO 11 I=1,NG
      D2(I)=0.
      NI=N(I)
      FNI=FLOAT(NI)
      FNIM1=FNI-1
      FNII=0.
      XBI=XB(I)
      DO 7 J=1,NI
      D2(I)=D2(I)+((FNI*X(J,I)-XBI)/FNIM1)**2
7     CONTINUE
      XBI=0.
      DO 8 II=1,NG
      IF(II.EQ.I)GOTO 8
```

```
      FNII=FNII+FLOAT(N(II))
      XBI=XBI+XB(II)
8     CONTINUE
      FNIIM1=FNII-1.
      DO 10 II=1,NG
      IF(II.EQ.I)GOTO 10
      NI=N(II)
      DO 9 J=1,NI
9     D2(I)=D2(I)+((FNII*X(J,II)-XBI)/FNIIM1)**2
10    CONTINUE
      D2(I)=D2(I)/FLOAT(NN)
11    CONTINUE
C COMPUTE THE THREE GROUP DISCREPANCY
      D3=0.
      DO 13 I=1,NG
      FNI=FLOAT(N(I))
      FNIM1=FNI-1.
      XBI=XB(I)
      DO 12 J=1,NI
12    D3=D3+((FNI*X(J,I)-XBI)/FNIM1)**2
13    CONTINUE
      D3=D3/FLOAT(NN)
C PRINT THE DISCREPANCIES
      WRITE(6,14)D1,D2,D3
14    FORMAT(5F13.1)
      STOP
      END
```

APPENDIX VII

```
∇ DEP ALLREG IND
 [1] ⍝THIS PROGRAM COMPUTES THE SELECTION CRITERIA FOR SUBSETS OF
 [2] ⍝REGRESSION VARIABLES UNDER THE HIGH STRUCTURE NORMAL MODEL
 [3] ⍝(SECTION 2.3) AND UNDER THE LOW STRUCTURE MODEL(SECTION 3.2)
 [4] NVAR←(⍴IND)[2]
 [5] ⍝THE NUMBER OF INDEPENDENT VARIABLES(INCLUDING THE ONE VECTOR
 [6] ⍝FOR THE MEAN)
 [7] TR←2*NVAR-1
 [8] ⍝TOTAL NUMBER OF POSSIBLE REGRESSIONS
 [9] INREG←(NVAR⍴2)⊤TR
 [10] ⍝INDICATOR VECTOR OF VARIABLES INCLUDED IN THIS PARTICULAR
 [11] ⍝REGRESSION; BEGINNING OF OUTER LOOP ON K
 [12] DK←0
 [13] LK←0
 [14] ⍝INITIALIZE SUMS
 [15] N←⍴DEP
 [16] ⍝NUMBER OF OBSERVATIONS
 [17] K←⍴(INREG=1)/INREG
 [18] ⍝NUMBER OF VARIABLES IN THIS REGRESSION
 [19] NK2←(N-K)÷2
 [20] A←N⍴1
 [21] A[1]←0
 [22] ⍝INDICATOR VECTOR OF OBSERVATIONS TO BE INCLUDED IN THIS
 [23] ⍝PARTICULAR REGRESSION
 [24] Z←INREG/[2]IND
 [25] ⍝THE INDEPENDENT VARIABLES MATRIX
 [26] ZJ←A/[1]Z
 [27] ⍝WITH THE JTH ROW OMITTED;BEGINNING OF THE INNER LOOP ON J
 [28] XJ←A/DEP
 [29] ⍝THE DEPENDENT VARIABLE WITH THE JTH OBSERVATION OMITTED
 [30] BJ←(⌹(⍉ZJ)+.×ZJ)+.×(⍉ZJ)+.×XJ
 [31] ⍝THE CORRESPONDING BETAHAT
 [32] ZJB←,(~A)/[1]Z
 [33] ⍝THE JTH ROW OF Z
 [34] CJ←1-ZJB+.×(⌹(⍉Z)+.×Z)+.×ZJB
 [35] D←(((~A)/DEP)-ZJB+.×BJ)*2
 [36] ⍝THE SQUARED DISCREPANCY FOR THE JTH OBSERVATION
 [37] DK←DK+D
 [38] WT←CJ÷+/(XJ-ZJ+.×BJ)*2
 [39] DWT←D×WT
 [40] LK←LK+(.5×⍟WT÷○1)-NK2×⍟(1+DWT)
 [41] A←1⌽A
 [42] →27×⍳A[1]≠0
 [43] LK←LK+⍟(!NK2-1)÷(!(N-K-3)÷2)
 [44] DK←DK÷N
 [45] 1↓INREG
 [46] DK
 [47] LK
 [48] ' '
 [49] TR←TR+1
 [50] →9×⍳TR<2*NVAR
∇
```